GSED

"Genetic Structures from Electrophoresis Data"

User's Manual

Elizabeth M. Gillet

Institut für Forstgenetik und Forstpflanzenzüchtung
Universität Göttingen
Büsgenweg 2 / 37077 Göttingen / Germany

April 1998

GSED User's Manual, April 1998

© E.M. Gillet, Institut für Forstgenetik, Univ. Göttingen, 1994, 1997, 1998

http://www.uni-forst.gwdg.de/forst/fg/software.htm

Contents

1 Introduction **1**

2 Constructing an input file **4**

2.1 Header 4

2.2 Format specification line 6

2.3 Input format for each sample 6

2.3.1 Key line 7

2.3.2 Genotypes of single individuals 7

2.3.3 Genotype frequencies in sample 8

2.3.4 End-of-sample line 9

2.4 End of input 9

2.5 Auxiliary programs 10

2.5.1 GSEDINPT 10

2.5.2 GSEDTEST 11

3 Running GSED **12**

3.1 First run 12

3.2 Configuration file 20

3.3 Sorting of haplotypes and genotypes 20

4 Understanding GSED output **21**

5 Frequency distributions **25**

6 Measures of variation **28**

6.1 Measures of variation within samples 28

6.1.1 Diversity v 28

6.1.2 Total population differentiation δ_T 28

6.1.3 Evenness e 28

6.2 Measures of variation between samples 29

6.2.1 Genetic distance d_0 29

6.2.2 Subpopulation differentiation D_j and δ 30
6.2.3 Test of homogeneity 30

7 Analysis of genotypic structure **31**
7.1 Heterozygosity 31
7.1.1 Proportion of heterozygosity of single-locus genotypes 31
7.1.2 Conditional heterozygosity of single-locus genotypes 31
7.1.3 Degree of heterozygosity of multilocus genotypes 31
7.2 Tests of single locus structure 32
7.2.1 Test of Hardy-Weinberg structure and heterozygosity 32
7.2.2 Test of product structure for ordered genotypes 33

8 Analysis of the gene pool **35**
8.1 Measures of variation within samples 35
8.1.1 Diversity v of the gene pool 35
8.1.2 Diversity v_{gam} of the hypothetical gametic output 35
8.1.3 Total population differentiation δ_T of the gene pool 36
8.2 Measures of variation between samples 37
8.2.1 Distance d_0 between gene pools 37
8.2.2 Differentiation δ of subdivided gene pools 37

9 Acknowledgements and disclaimer **38**

A Technical considerations **41**
A.1 Hardware requirements 41
A.2 Limitations on data 41
A.3 Temporary direct access storage 41

B Examples of input files **46**

List of Tables

1 *Frequency distributions calculated by GSED* 1

2 *Characterization of genetic structures* 2

3 *Examples of format specification lines.* 5

4 *Examples of key lines using the format specification "*`10 (2I4,1X,10(2I2))`*"*. 8

5 *Example of a skeleton input file constructed by the auxiliary program GSEDINPT. Using any (ASCII) text editor, the lines beginning with "*`<-`*" must be replaced by the genotypic data for the respective population sample, right-justified within the fields defined by the key line. Note that gametic sex is specified for all loci in Population 1, two loci in Population 2, and no loci in Population 3.* 10

6 *Interactive sequence for first run of GSED for an input file EXAMPLE.DAT (continued in next table).* 13

7 *Continuation of the interactive sequence begun in the previous table.* . . . 14

8 *Start of the interactive sequence for a subsequent run using the input file EXAMPLE.DAT. The choices made during the first run were stored in the configuration file EXAMPLE.CFG shown below. This configuration can be adopted by replying with a "*`Y`*", in which case the interactive sequence will be skipped. A reply of "*`N`*" allows a new choice of frequency distributions and calculations.* 17

9 *Examples demonstrating designation of alleles, haplotypes and genotypes in the output.* 23

10 *Example of output for a test of genotypic structure.* 23

11 *Interpretation and maximal values of variables used in calculating storage space needed for direct access files* 42

12 *Maximum size of the direct access files as number of 4-byte integers. Multiplication of each result by 4 yields total number of bytes. Notation: (1) If additional storage space is added blockwise, add Size.of.Enlargement. (2) Max.No.Haplotypes = Max.No.Alleles if only single loci are considered.* 43

1 Introduction

The computer program GSED ("Genetic Structures from Electrophoresis Data") is based on a conceptually and mathematically unified system of data analysis for the characterization of genetic structures in population genetic investigations, which has been developed at this institute in recent years. Tabs. 1 and 2 contain an overview.

All measures of variation within and between samples and gene pools are based on the single measure of genetic distance d_0, which quantifies the proportion of genetic types by which two collections of individuals or populations differ ([4], [5], [6], [8]). Heterozygosity goes even deeper by measuring variation at the level of the individual. Inference on the manner of association of alleles to form genotypes is provided by tests of two models describing mating systems.

The input to GSED consists of the genotypes found in a sample of individuals taken from a collection (e.g., population, deme, cohort). An individual's **genotype** refers to the alleles present at a single gene locus (**single locus genotype**) or, more commonly, at each of a number of gene loci (**multilocus genotype**). If known, designation of the gametic sex of each allele at a locus, that is, the sex of the contributing parent, can be included in the genotype.

Many of the measures of variation calculated by the program can be applied not only to genetic types but also to any system of classification by which each individual of a population can be assigned one of a finite set of discrete "types" (e.g., phenotypes, ecotypes). Although the assumption that data input to GSED concerns genetic types is reflected in its commentaries, one or higher dimensional non-genetic classifications can be disguised as maternal alleles or haplotypes at "loci" for which gametic sex is specified and "paternal" type unknown. An input file would be analogous to that of Example 4 in B. Output headings would have to be reinterpreted accordingly.

When using GSED to analyze genetic types, however, it is essential that the alleles at each locus be known. In other words, the phenotype produced by the genes at each locus must be a **gene marker**, in that the phenotype enables identification of *all* involved

Table 1: *Frequency distributions calculated by GSED*

Single locus	Allele frequencies among maternal contributions*
	Allele frequencies among paternal contributions*
	Allele frequencies
	Genotype frequencies
Multilocus	Haplotype frequencies among maternal contributions*
	Haplotype frequencies among paternal contributions*
	Haplotype frequencies*
	Genotype frequencies

* If gametic sex of the alleles at each locus is specified

Table 2: *Characterization of genetic structures*

- Analysis of allelic, haplotypic and genotypic structures
  - *Measures of variation within samples*
    - Diversity v
    - Total population differentiation δ_T
    - Evenness e
  - *Measures of variation between samples*
    - Genetic distance d_0
    - Subpopulation differentiation D_j and δ
    - Test of homogeneity
- Analysis of genotypic structure
  - Heterozygosity, single locus and multilocus
  - Test of Hardy-Weinberg structure and heterozygosity
  - Test of product structure
- Analysis of the gene pool
  - *Measures of variation within samples*
    - Diversity v of the gene pool
    - Diversity v_{gam} of the hypothetical gametic output
    - Total population differentiation δ_T of the gene pool
    - *Measures of variation between populations*
    - Distance d_0 between gene pools
    - Differentiation δ of subdivided gene pools

alleles. Isoenzyme phenotypes resulting from gene loci possessing a codominant mode of inheritance are gene markers. Dominance, such as is caused by a (recessive) null allele at a locus in isoenzyme investigations, does not give rise to a gene marker; loci showing dominance cannot be used for the analysis of genetic types unless additional inheritance analysis has revealed the true genotype of each individual at the locus.

GSED is written in FORTRAN-77 and has been compiled for the operating systems MS-DOS/PC-DOS (Version 2.1 or higher), VMS (Version 5.0 or higher), and TOS (Atari). The program reads data from a previously constructed ASCII input file. In each run, the user interactively requests calculations of frequency distributions, measures and tests, and samples. The output is printed either directly onto the screen or stored in a file for later processing.

The next three sections (2, 3, 4) of this manual deal with practical matters, namely constructing an input file, running the program, and understanding the output. The following four sections (5, 6, 7, 8) are concerned with the concepts behind the program. The first of these sections (5) reviews the different types of frequency distribution that can be calculated from samples of multilocus genotypes. Three sections (6, 7, 8) outline the measures and tests that are performed for the various frequency distributions, including references to (mostly original) articles containing detailed descriptions of the underlying concepts. A list of references follows, denoted in the text by numbers in square brackets. Two appendices (A, B) are devoted to technical considerations (hardware, storage space) and examples of input files.

2 Constructing an input file

The input to GSED consists of the genotypes of the individuals included in samples taken from one or more populations (demes, provenances, etc.). GSED reads the genotypes represented in each sample from an external file that was constructed beforehand using any (ASCII) text editor.

Each input file for GSED consists of three parts, each of which will be explained in detail subsequently:

- a header defining the names and numbers of gene loci and samples (see Sec. 2.1);
- a format specification line giving both the number of loci per genotype and a FORTRAN format specification defining the character positions in which data appear in subsequent lines (see Sec. 2.2);
- the single locus or multilocus genotypes observed in each sample. Either the genotype is given for each single individual in the sample (see Sec. 2.3.2), or each genotype found in the sample is accompanied by its frequency, *i.e.*, the number of individuals possessing it (see Sec. 2.3.3).

Two auxiliary programs aid input file construction (see Sec. 2.5):

- GSEDINPT writes header and key lines to a file according to information interactively supplied by the user (see Sec. 2.5.1).
- GSEDTEST tests finished input files by reading them and typing them on the screen (see Sec. 2.5.2).

Examples of input files are given in B.

2.1 Header

The header of an input file has the following form:

- One line containing number of samples (maximum of 100) and number of gene loci (between 1 and 50), separated by a comma or blank;
- Gene locus names (maximum of 12 characters each) on consecutive lines;
- Sample names (maximum of 40 characters each) on consecutive lines.

Examples of format specification lines are given in Tab. 3.

Table 3: *Examples of format specification lines.*

The format specification line

```
10  (2I4,1X,10(2I2))
.........+.........+
```

reads 22 integers, including a 10-locus genotype, from the following line of data (namely 2, 123, 1, 2, 2, 3, .., 3, 3):

```
   2 123  1 2  2 3  3 5  2 1  2 3  2 1  1 1  1 1  2 2  3 3
.........+........+.........+.........+.........+.........+
```

The format specification line

```
20  (I4,I4,10(2I2)/9X,10(2I2))
.........+.........+.........+
```

reads 42 integers, including a 20-locus genotype, from the following two lines of data (2, 2345, 2, 1, 2, 3, .., 1, 1):

```
   22345  21 23 24 22 33 11 21 21 23 21
          32 33 12 23 23 31 32 23 32 11
.........+........+.........+.........+
```

2.2 Format specification line

- The **format specification line** contains
  - the number of loci constituting the (multilocus) genotypes in character positions 1–2, and
  - a FORTRAN format specification beginning in character position 5, defining which character positions of the subsequent data lines contain the information required by the program.

As explained in 2.3, the FORTRAN format specification must specify the character positions of $2 + 2 * n$ integers in each of the subsequent input records, where n is the number of gene loci in the multilocus genotypes.

For present purposes, a general description of the following elements of a FORTRAN format specification will suffice (for further information consult any FORTRAN language reference manual):

`nIw`	The `I` field descriptor indicates that `n` integers are to be read in consecutive fields of width `w` (*i.e.*, number of character positions)
`nX`	The `X` field descriptor indicates that `n` character positions are to be skipped
`r(...)`	The repeat count `r` indicates that the contents of the parentheses are to be repeated `r` times
`,`	Separates field descriptors
`/`	Separates field descriptors and causes reading to continue on a new line
`(...)`	Parentheses enclose the entire format specification

2.3 Input format for each sample

The main part of an input file consists of the genotypes found in each sample.

The data for each sample has the following form:

- A key line describing how the following genotypes are to be interpreted;
- The list of genotypes found in the sample;
- End-of-sample line consisting of the integer "`9999`" in the first `I`-field.

The end-of-sample line for one sample is followed immediately (no empty lines) by the key line of the next sample. The samples can appear in any order. The input file terminates with the end-of-sample line of the last sample.

All data lines, including key line and end-of-sample line, are read according to the format specification line described above (see Sec. 2.2).

2.3.1 Key line

The **key line** of a sample fulfills four functions:

- designation of the sample number;
- assignment of fields (*i.e.*, blocks of character positions) in a data line to gene loci;
- specification of whether each of the genotypes is to be interpreted as that of a single individual (see Sec. 2.3.2) or of a number of individuals (see Sec. 2.3.3);
- indication of whether gametic sex is specified for the alleles at each locus (see Sec. 5).

The key line specifies the following integers for n gene loci:

sampleno *keyline*="0" $\pm$*locdef*$_1$ *gamsex*$_1$ *locdef*$_2$ *gamsex*$_2$ ... *locdef*$_n$ *gamsex*$_n$

where

sampleno	sample number, referring to list of samples in header
keyline="0"	indicator of key line
locdef$_1$	locus number of first locus in multilocus genotype, referring to list of gene loci in header (usually ="1")
+*locdef*$_1$ or *locdef*$_1$	indicates that each multilocus genotype is that of a single individual (see Sec. 2.3.2)
$-$*locdef*$_1$	indicates that each multilocus genotype is accompanied by its frequency in the sample (see Sec. 2.3.3)
gamsex$_1$	indicator of gametic sex specification of first locus: *gamsex*$_1$="1" if gametic sex is specified, "0" otherwise
locdef$_i$	locus number of ith locus in multilocus genotype,
gamsex$_i$($i = 2, \ldots, n$)	indicator of gametic sex specification of ith locus: *gamsex*$_i$="1" if gametic sex is specified, "0" otherwise

Examples of key lines are given in Tab. 4.

2.3.2 Genotypes of single individuals

If the sign of *locdef*$_1$ in the key line of a sample equals "+" or is blank (see Sec. 2.3.1), then each genotype is interpreted to be that of a single individual. In this case, each genotype line specifies the following integers:

sampleno *indivno* *locus*$_1$ *allele*$_1$ *locus*$_1$ *allele*$_2$ ... *locus*$_n$ *allele*$_1$ *locus*$_n$ *allele*$_2$

where

Table 4: *Examples of key lines using the format specification "*`10  (2I4,1X,10(2I2))`*".*

Key line for sample 2 specifying that multilocus genotypes comprise loci 1–10, that each genotype is that of a single individual, and that gametic sex is not specified for any locus:

```
   2   0  1 0  2 0  3 0  4 0  5 0  6 0  7 0  8 0  9 0 10 0
.........+........+.........+.........+.........+.........+
```

Key line for sample 5 specifying that multilocus genotypes comprise loci 1–10, that each genotype is accompanied by its frequency in the sample, and that gametic sex is specified for all loci:

```
   5   0 -1 1  2 1  3 1  4 1  5 1  6 1  7 1  8 1  9 1 10 1
.........+........+.........+.........+.........+.........+
```

sampleno	sample number, referring to list of samples in header
indivno	number designating the individual whose genotype is listed
$locus_i\, allele_1$	designation of first allele at locus i as an integer $\geq$"`-1`"; if gametic sex is specified for this locus, then $locus_i\, allele_1$ is the allele contributed by the maternal parent
$locus_i\, allele_2 (i = 2, \ldots, n)$	designation of second allele at locus i as an integer $\geq$"`-1`"; if gametic sex is specified for this locus, then $locus_i\, allele_2$ is the allele contributed by the paternal parent

2.3.3 Genotype frequencies in sample

If the sign of $locdef_1$ in the key line of a sample equals "`-`" (see Sec. 2.3.1), then each genotype is interpreted as having been found in a number of individuals. The frequency of a genotype in the sample is given in the second field. In this case, each genotype line specifies the following integers:

sampleno *frequency* $locus_1\, allele_1$ $locus_1\, allele_2$ ... $locus_n\, allele_1$ $locus_n\, allele_2$

where

sampleno	sample number, referring to list of samples in header
frequency	number of individuals possessing the genotype
$locus_i\, allele_1$	designation of first allele at locus i as an integer $\geq$"`-1`"; if gametic sex is specified for this locus, then $locus_i\, allele_1$ stems from the **maternal parent**
$locus_i\, allele_2 (i = 2, \ldots, n)$	designation of second allele at locus i as an integer $\geq$"`-1`"; if gametic sex is specified for this locus, then $locus_i\, allele_2$ stems from the **paternal parent**

A **null allele** at locus i is designated by $locus_i\, allele_j$="`0`" (zero, $j = 1$ or 2).

An **unknown allele** is specified by $locus_i\, allele_j$="`-1`" ($j = 1$ or 2). Note that unknown alleles in a random sample of genotypes can present a problem for the calculation of frequency distributions (see Sec. 5).

2.3.4 End-of-sample line

The end-of-sample line contains only the following integer in the first field:

sampend = "`9999`"

Data for the remaining samples are appended to the previous end-of-sample line according to the same pattern.

Since GSED does not know when to expect an end-of-sample line, when encountered, these lines are read according to the same format specification (see Sec. 2.2) as the key and genotype lines. Some compilers specify that data be read in all of the fields defined in the format specification, even if a line is terminated after the first field; they seek the "missing" data in the next lines (which probably already define the next sample) and print an error message as soon as the mix-up leads to ambiguity. One solution is to fill out all of the remaining fields defined by the format specification with blanks, including those in continuation lines. As this is often forgotten, GSED uses the following strategy: If "`9999`" is encountered in field 1, the current sample is ended. The program then backspaces in the input file to this same end-of-sample line and begins to search line for line for the next key line (recognizable by the "`0`" in field 2). In order for this to work, the sample number and the "`0`" in the first two fields of the next key line must be separated either by one or more blank characters or by a comma. By this means reading errors should be avoided.

2.4 End of input

Reading of data is terminated when the end of the input file is encountered, that is, after the end-of-sample line for the final sample.

Operating systems apparently differ in their handling of end-of-files: Whereas TOS recognizes the end of an input file if no further characters appear after the last "`9999`" (not even "`<Return>`" is necessary), DOS expects to find the cursor at the beginning of the

Table 5: *Example of a skeleton input file constructed by the auxiliary program GSEDINPT. Using any (ASCII) text editor, the lines beginning with "<-" must be replaced by the genotypic data for the respective population sample, right-justified within the fields defined by the key line. Note that gametic sex is specified for all loci in Population 1, two loci in Population 2, and no loci in Population 3.*

```
.........+.........+.........+.........+.........+.........+.........+.......
   3   4
LAP-A
LAP-B
IDH
PGI
Population 1
Population 2
Population 3
 4  (2I4, 4(1X,2I2))
   1   0  1 1  2 1  3 1  4 1
 <- Enter sample no., individual no., and genotype of all sampled individuals
 <- listing maternal allele at each locus first
9999
   2   0  1 1  2 1  3 0  4 0
 <- Enter sample no., individual no., and genotype of all sampled individuals
 <- listing maternal allele first if gametic sex is specified
9999
   3   0 -1 0  2 0  3 0  4 0
 <- Enter sample no., frequency, and genotype for all sampled genotypes
9999
.........+.........+.........+.........+.........+.........+.........+.......
```

next (empty) line. To compensate for these differences, an arbitrary number of empty lines can appear after the final "9999" — for DOS, at least one such line is mandatory!

2.5 Auxiliary programs

2.5.1 GSEDINPT

GSEDINPT is a self-explanatory, interactive program with the help of which a skeleton input file can be constructed. In accordance with the user's answers to questions about the data, the following lines are written into the new file: header, format specification line, and for each sample the key line and the end-of-sample line.

Tab. 5 gives an example of a skeleton file (EXAMPLE.DAT) constructed by GSEDINPT.

After completion of GSEDINPT, the genotype data for each sample must be inserted between the respective key line and the end-of-sample line with the help of any (ASCII)

text editor. Positioning of the genotype data over the character positions of a line obeys the example given by the key line, *i.e.*, the integers are placed right-justified within the fields defined by the key line. Examples of completed input files are given in B.

2.5.2 GSEDTEST

GSEDTEST is a self-explanatory program to quickly test new or altered input files for their readability by GSED. GSEDTEST reads the lines in an input file as would GSED and types them on the screen or into a file. An error message is printed if an unreadable data line is encountered.

3 Running GSED

Execution of GSED begins interactively with a sequence of questions to be answered by the user. After specifying a preconstructed input file (see Sec. 2), the user may choose any or all of the frequency distributions listed in Tab. 1 and request calculation of any of the measures and tests listed in Tab. 2. Additional questions concern the format of the output. After all questions have been answered, GSED performs the desired calculations. Results can either be written into an output file or typed on the screen.

A sample interactive sequence is given in Tabs. 6 and 7.

3.1 First run

The first time an input file is read into GSED, the sequence of questions listed in Tabs. 6 and 7 appears on the screen. The meaning of these questions will be described in more detail in the following. In cases where choice of option is specified by a capital letter, small letters are also accepted (e.g., "`y`" instead of "`Y`").

```
.....................................................................
Input file ? :
```

Type the name of the input file (see Sec. 2), including path specification if necessary. A maximum of 60 characters is allowed.

```
.....................................................................
Locus configuration ?     "0" : all single loci    "2" : multilocus - some loci
                          "1" : some single loci   "3" : multilocus - all loci
Option ? :
```

The four options are explained as follows:

```
Option ? : 0
```

Calculations will be carried out for every single locus.

```
Option ? : 1
```

Calculations will be carried out for some of the single loci. As an example, the loci 2 and 4 are specified in reply to the following question:

Table 6: *Interactive sequence for first run of GSED for an input file EXAMPLE.DAT (continued in next table).*

(GSED logo)

```
Input file ? : EXAMPLE.DAT
.....................................................................
Locus  1: LAP-A
Locus  2: LAP-B            Sample  1: Population 1
Locus  3: IDH              Sample  2: Population 2
Locus  4: PGI              Sample  3: Population 3
.....................................................................
Locus configuration ?     "0" : all single loci    "2" : multilocus - some loci
                          "1" : some single loci   "3" : multilocus - all loci
Option ? : 0
.....................................................................
Choice of frequency distributions (Answer "Y" (yes) or "N" (no))
   Allele/haplotype frequencies among maternal contributions ? : Y
   Allele/haplotype frequencies among paternal contributions ? : Y
   Allele/haplotype frequencies ? : Y
   Genotype frequencies ? : Y
.....................................................................
Should gametic sex specification, if given, be retained ? : Y
.....................................................................
Choice of calculations (Answer "Y" (yes) or "N" (no))

   Frequency distributions ? : Y

   Measures of variation within samples
   ------------------------------------
   Diversity v_2 ? : Y
   Total population differentiation  delta_T? : Y
   Evenness : Y
     -    finite population size ? : Y
     -  infinite population size ? : Y

   Measures of variation between samples
   -------------------------------------
   Genetic distance d_0 ? : Y
   Subpopulation differentiation D_j, delta : Y
     -  subpopulations weighted proportional to sample size ? : Y
     -  subpopulations equally weighted ? : Y
   Test of homogeneity of the sample distributions ? : Y

   Analysis of genotypic structure
   -------------------------------
   Heterozygosity ? : Y
   Tests of single locus structure :
      Test of Hardy-Weinberg structure and heterozygosity ? : Y
      Test of product structure (only if gametic sex is specified) ? : Y
```

Table 7: *Continuation of the interactive sequence begun in the previous table.*

```
Should these choices be stored in a file for later use ? Y
Configuration stored in file EXAMPLE.CFG
.....................................................................
Sample  1: Population 1
Sample  2: Population 2
Sample  3: Population 3

Samples for output ?      "0" : all samples        "1" : some samples
Option ? : 0
.....................................................................
Output unit ?             "S" : screen             "F" : file
Option ? : F

Output file ? : EXAMPLE.OUT
.....................................................................
Width of output (min. of 75 characters/line) as number of samples per line ? :
     No. samples/line = 1/10 * (No.characters/line - 15)
For example:
     "0"  for ALL   3 samples ( 75 char/line)
     "6"  for       6 samples ( 75 chars/line as for DIN A4 paper upright)
     "10" for      10 samples (115 chars/line as for DIN A4 paper crosswise)
     "11" for      11 samples (125 chars/line as in condensed mode)
Option ? : 0
.....................................................................
Reading input file EXAMPLE.DAT
   - sample    1: Population 1
                  (     7 records found)
   - sample    2: Population 2
                  (     7 records found)
   - sample    3: Population 3
                  (    11 records found)
Reading of input file completed.

-- Sorting haplotypes ...
-- Sorting genotypes ...
-- Calculating and outputting results ...
   - for locus No.    1:  LAP-A
   - for locus No.    2:  LAP-B
   - for locus No.    3:  IDH
   - for locus No.    4:  PGI
.....................................................................
Additional calculations using the same input file and locus configuration ? :
Option ? : N
FORTRAN STOP
```

```
Number of different single loci ? : 2
Which gene loci (separated by commas and using as many lines as necessary) ? :
2,4
```

```
Option ? : 2
```

Calculations will be carried out for multilocus genotypes defined by the genotypes at different sets of gene loci, the so-called **multilocus combinations**. As an example, one multilocus combination comprising the gene loci 1 and 2 and a second comprising only the single locus 1 are specified in reply to the following questions (as in the second case, a "multi"locus combination can refer to the genotypes at a single locus):

```
Number of different multilocus combinations ? : 2
Combination 1 :  How many gene loci ? 2
   Which loci (separated by commas and using as many lines as necessary) ? :
1,2
Combination 2 :  How many gene loci ? 1
   Which loci (separated by commas and using as many lines as necessary) ? :
3
```

It is important to note here that measures characterizing the gene pool (see Sec. 8) are calculated if and only if the "`Locus configuration ?`" comprises only single loci, *i.e.*, if either option "`0`" or "`1`" is chosen. Since the gene pool measures are formulated as means of the respective single-locus measures at all loci contributing to the gene pool, the single-locus measures must already be available.

```
Option ? : 3
```

Calculations will be performed for multilocus genotypes defined by the genotypes at all of the single loci (in EXAMPLE.DAT for all four loci).

```
.....................................................................
Choice of frequency distributions (Answer "Y" (yes) or "N" (no))
```

Choices can be made among the four types of frequency distribution offered by the subsequent questions and described below (see Sec. 5).

```
.....................................................................
Should gametic sex specification, if given, be retained ? : Y
```

If no gametic sex is specified at any locus, then the answer to this question is meaningless.

If gametic sex is specified at some or all loci, then an answer of "N" will cause this specification to be ignored at all of them. For example, in such a case both of the genotypes A_1A_2 and A_2A_1, where the first allele is that contributed by the maternal parent, would be counted as the genotype A_1A_2.

```
.....................................................................
Choice of calculations (Answer "Y" (yes) or "N" (no))
Frequency distributions ? : Y
```

An answer of "Y" causes the calculated frequency distributions to be included in the output. If the answer is "N", they will be omitted.

```
Measures of variation within samples . . .
Measures of variation between samples . . .
Analysis of genotypic structure . . .
```

If selected, the measures and tests offered in the subsequent questions (and described in Sec. 6, 7, 8) are calculated for each of the chosen frequency distributions.

```
Should these choices be stored in a file for later use ? Y
```

If the answer is "Y", the above choices of frequency distributions, measures, and tests are saved in a configuration file. The configuration file has the same filename as the input file and the extension ".CFG". All subsequent runs using the same input file will first print the stored configuration table and then ask whether it should be adopted. If the answer is "N", new choices can be made.

An example for the case in which a configuration file already exists is given in Tab. 8.

The sequence of questions continues as follows:

```
.....................................................................
Samples for output ?      "0" : all samples        "1" : some samples
Option ? :
```

The two options are explained as follows:

```
Option ? : 0
```

The output contains the results for all of the samples in the input file. Measures of variation between samples (see 6.2) are calculated using ALL of the samples.

Table 8: *Start of the interactive sequence for a subsequent run using the input file EXAMPLE.DAT. The choices made during the first run were stored in the configuration file EXAMPLE.CFG shown below. This configuration can be adopted by replying with a "*`Y`*", in which case the interactive sequence will be skipped. A reply of "*`N`*" allows a new choice of frequency distributions and calculations.*

```
Input file ? :
EXAMPLE.DAT
.....................................................................
Input file : EXAMPLE.DAT

Locus  1: LAP-A
Locus  2: LAP-B            Sample  1: Population 1
Locus  3: IDH              Sample  2: Population 2
Locus  4: PGI              Sample  3: Population 3
.....................................................................
Locus configuration ?     "0" : all single loci    "2" : multilocus - some loci
                          "1" : some single loci   "3" : multilocus - all loci
Option ? : 0
.....................................................................
Configuration file exists : EXAMPLE.CFG
-----------------------------------------------------------------------------
| Frequency distributions                   Y | Maternal contributions       Y|
| ....MEASURES.OF.VARIATION.WITHIN.SAMPLES... | Paternal contributions       Y|
| Diversity                                 Y | Allele/haplotype frequencies Y|
| Total population differentiation deltaT   Y | Genotype frequencies         Y|
| Evenness                                  Y ---------------------------------|
|  - finite population size                 Y                                  |
|  - infinite population size               Y                                  |
| ...MEASURES.OF.VARIATION.BETWEEN.SAMPLES...                                  |
| Genetic distance                          Y                                  |
| Subpopulation differentiation D_j, delta  Y                                  |
|  - weights proportional to sample size    Y                                  |
|  - equal weights                          Y                                  |
| Test of homogeneity                       Y                                  |
| .................ANALYSIS.OF.GENOTYPIC.STRUCTURE.................          |
| Heterozygosity                                                  Y            |
| Test of Hardy-Weinberg structure + heterozygosity (single locus)   Y         |
| Test of product structure                        (single locus)   Y          |
-----------------------------------------------------------------------------
Do you want to adopt this configuration ? Answer "Y" (yes) or "N" (no): Y
```

```
Option ? : 1
```

The output contains the results for only those samples given in reply to the following question. This option allows measures of variation between samples (see Sec. 6.2) to be calculated for differing sets of samples. As an example, samples 1 and 3 are chosen in reply to the following questions:

```
How many samples ? : 2
   Which samples (separated by commas and using as many lines as necessary) ? :
1,3



.....................................................................
Output unit ?             "S" : screen             "F" : file
Option ? :
```

Output can be directed to one of two units as follows:

```
Option ? : S
```

All results are typed on the screen. They are not saved elsewhere and thus are lost as soon as they disappear off the screen.

```
Option ? : F
Output file ? : EXAMPLE.OUT
```

Results are output as ASCII text to the designated file. A maximum of 60 characters are allowed for the file name and any necessary specification of path. Since the output is in ASCII code, it is possible to alter its format later using any text editor. The finished file can then be printed on any printer.

If the output file EXAMPLE.OUT already exists, the following message appears:

```
File already exists:      "A" : append new output  "O" : overwrite old output
Option ? :



Option ? : A
```

The new output will be added to the end of the existing file, thus preserving the previous contents.

```
Option ? : O
```

The new output will replace the previous contents of the file, which are thus lost. Note that this option is indicated by the letter "O" and not the numeric character "0" (zero).

```
.....................................................................
Width of output (min. of 75 characters/line) as number of samples per line ? :
     No. samples/line = 1/10 * (No.characters/line - 15)
For example:
     "0"  for ALL   3 samples ( 75 char/line)
     "6"  for       6 samples ( 75 chars/line as for DIN A4 paper upright)
     "10" for      10 samples (115 chars/line as for DIN A4 paper crosswise)
     "11" for      11 samples (125 chars/line as in condensed mode)
Option ? : 0
```

The width of the output medium (e.g., paper) can vary and with it the number of samples that fit onto one line. If the available number of characters per line is known, the formula on the second line above yields the maximal number of samples (rounding down to the nearest integer, if necessary). If not all of the samples fit onto one line, the tables of results are cut off after the specified number of samples and continued on the next lines of output.

The minimum number of characters per line is set to 75. One reason is that this is the length of the commentaries in the output. Another is that the maximal number of columns of the contingency tables in the tests of genotypic structure that can be printed onto one line is also set to the chosen number of samples per line.

A reply of "0" (zero) causes the results of all samples to be printed onto one line.

```
.....................................................................
Additional calculations using the same input file and locus configuration ? :
Option ? :
```

When this line appears, the chosen calculations have been completed and either typed on the screen or stored in the output file. Its purpose becomes apparent in the description of the options:

```
Option ? : Y
```

Type "Y" if additional frequency distributions, measures or tests for the same or a different set of samples are desired for the same input file and locus configuration. Since the frequency data is already stored, the input file is not reread and results are obtained quickly. This option provides a means of ordering the output differently from that reflected by the interactive sequence. It also allows calculation of measures of variation between samples for different sets of samples.

```
Option ? : N
```

An answer of "N" terminates the program.

3.2 Configuration file

In subsequent runs of GSED for an input file, a configuration file may exist. This will be the case, if the question "Should these choices be stored in a file for later use ?" was answered with "Y" in an earlier run for the same input file. The configuration file contains the previous answers to the questions listed under the headings of "Choice of frequency distributions" and "Choice of calculations" (see Tab. 6). Its name is composed of the filename of the input file and the extension .CFG.

If a configuration file exists, then a configuration table such as that presented in Tab. 8 is typed on the screen after the "Locus configuration" has been specified (see Tab. 6).

If the answer to the subsequent question "Do you want to adopt this configuration ?" is answered by "Y", then "Choice of frequency distributions" and "Choice of calculations" are skipped. The question "Should gametic sex specification, if given, be retained ?" is still posed, since in the case of gametic sex specification, one may want to perform the same calculations with and without regard of gametic sex (see Sec. 5).

If the answer to "Do you want to adopt this configuration ?" is "N", then "Choice of frequency distributions" and "Choice of calculations" must be made anew, as in Tab. 6.

3.3 Sorting of haplotypes and genotypes

An answer of "Y" to any of the following questions causes the lists of encountered haplotypes and genotypes to be printed in lexicographic order:

```
Frequency distributions ? :
Test of homogeneity of the sample distributions ? :
Test of Hardy-Weinberg structure and heterozygosity ? :
Test of product structure (only if gametic sex is specified) ? :
```

Since sorting of multilocus types can take an extreme amount of computing time, it is advisable not to choose these calculations for multilocus combinations (only the first two questions apply) unless they themselves are of interest. A test of homogeneity for a large number of multilocus types may well exceed the capacity of the program anyway (see Sec. A.2). Often, heterozygosity is the only calculation desired for multilocus genotypes; it is performed quickly if it alone is selected.

4 Understanding GSED output

GSED output is divided into the output for each locus combination, *i.e.*, for each single locus or multilocus combination (see Sec. 6, 7). If calculations are requested only for single loci, results for the gene pool and hypothetical gametic output defined by these loci are included at the close of the output (see Sec. 8).

Each locus combination (single or multilocus) is in turn divided into the output for each of the chosen frequency distributions (measures of variation within samples, between samples) followed by the analysis of genotypic structure (heterozygosity, tests of single locus structure).

Results for measures of variation and heterozygosity appear in tables, each column containing the results for one of the chosen samples. If the chosen width of output (see Sec. 3.1: "`Width of output ...`") is not sufficient to allow inclusion of all samples onto one line, each table is truncated vertically and continued on the next lines. If the current locus combination consists only of a single locus, the output for this combination closes with the results of the chosen tests of single locus structure for each sample.

The output for each frequency distribution begins with a heading which provides the following information about each sample:

"`Sample No.`"	Number of the sample in accordance with the list of samples in the input file.
"`Gam.sex.spec.?`"	Abbreviation of "Gametic sex specification?" = "`yes`", if the sex of the parent contributing each allele is known in the entire sample; = "`no`", otherwise.
"`Sample size`"	Total number of individuals whose genotypes are included in the input file, regardless of whether they contain unknown alleles or not.
"`No. identified`"	Number of individuals whose genetic types with respect to the current frequency distribution are completely identified (no unknown alleles, see Sec. 5). Relative frequencies refer to this number.
"`alpha`"	see below
"`alpha - HWP`"	see below
"`No. unknown`"	Number of individuals whose genetic types with respect to the current frequency distribution are unknown and thus are not counted (see Sec. 5).

"`Sample size`" equals the sum of "`No. identified`" and "`No. unknown`".

"`alpha`": In loose terms, "`alpha`" tells how frequent a type (allele, haplotype, genotype) must be in the base population in order for it to have a probability of 0.95 or greater of being represented in a sample of the given size ("`No. identified`"). More precisely, the

probability of having sampled (and identified) all types occurring with relative frequency greater than or equal to "alpha" is 0.95 or greater. Obviously, the larger the sample size, the smaller "alpha" becomes. (see [7] for derivation of "alpha".)

Alleles and multilocus haplotypes occur in pairs in the form of genotypes. If the only way to sample haplotypes is by sampling genotypes, it must be remembered that the manner of association between the different haplotypes making up the genotypes (homozygosity, heterozygosity) has a great influence on the probability of finding the rarer haplotypes. "alpha" describes the worst-case situation for finding the rarer haplotypes, namely pure homozygosity, in that only one allele or haplotype can be sampled per individual (see [7] for proof.)

"alpha - HWP": In the case of alleles and haplotypes with arbitrary frequencies, this relative frequency characterizes an analogous "alpha" for the best-case situation for sampling haplotypes when only genotypes can be sampled. [7] gives proof that this situation occurs when the genotypes arose by random fertilization between alleles/haplotypes, which are then independently associated in the genotypes. The resulting Hardy-Weinberg-Proportions (HWP) thus represent the optimal relationships between homozygosity and heterozygosity for sampling different alleles or haplotypes in genotypes. In [7] it is shown that "alpha - HWP" is equal to the value of "alpha" for a sample twice the size of the given sample. Thus sampling haplotypes in a Hardy-Weinberg population of genotypes is equivalent to drawing a sample of haplotypes singly (as opposed to pairwise) that is twice the size of the given sample of genotypes.

The output for the various measures and tests is described in the relevant Sec. 6, 7, 8.

The legend printed at the beginning of the output explains a few additional conventions:

```
  ( )        observed relative frequency
  [ ]        expected absolute frequency in a test
  { }        multilocus haplotype or genotype
"-9.999"     "undefined"
```

The designation of the different alleles, haplotypes, and genotypes in the output is demonstrated in Tab. 9.

The output of the statistical tests is similar to the example in Tab. 10.

The upper table in 10 contains the observed frequencies of the genotypes A_1A_1, A_1A_3, and A_3A_3 and, beneath each in square brackets, the frequencies expected under the null hypothesis of Hardy-Weinberg structure. The observed allele frequencies are given to the right of the table. The lower table, entitled "Test statistics", contains the results of the likelihood ratio test ("G"), Pearson's χ^2 test ("X"), and, in tables such as this with one degree of freedom ("DF=1"), the χ^2 test with continuity correction of 0.5 ("X**2(c=.5)"). The symbol (here "*") directly to the right of each statistic indicates its level of significance, which can be inferred from the two rightmost columns of the table: The abbreviation "C.V. of CHI**2" stands for "critical value of the χ^2 distribution" for

Table 9: *Examples demonstrating designation of alleles, haplotypes and genotypes in the output.*

Genetic type	Designation	Consists of
Allele	`1`	allele 1
Single-locus genotype	`1 x 3`	alleles `1` and `3`
Ordered single-locus genotype	`3 x 1`	maternal allele `3` × paternal allele `1`
Haplotype	{ 2 4 1 }	allele `2` at first locus + allele `4` at second locus + allele `1` at third locus
Multilocus genotype	{ 1 4 2 3 1 3 }	single-locus genotypes `1 x 4` at first locus + `2 x 3` at second locus + `1 x 3` at third locus
Ordered multilocus genotype	{ 1 2 4 0 3 2 }	maternal haplotype { 1 4 3 } × paternal haplotype { 2 0 2 }

Table 10: *Example of output for a test of genotypic structure.*

```
---- TEST OF HARDY-WEINBERG STRUCTURE (GAMETIC SEX DISREGARDED) ----

Allele      1         3    Freq.
       --------------------
    1  |   10             |   80
       [   16.00]         |
    3  |   60        30   |  120
       [   48.00][   36.00]
       -------------------------
                          |  200

                                    Level of   C.V. of CHI**2
        Test statistics            significance    (DF=  1)
    ------------------------------------------------------------
    G          =    6.438 *         0.050(*  )       3.841
    X**2       =    6.250 *         0.010(** )       6.635
    X**2(c=.5) =    5.486 *         0.001(***)      10.828
```

the given degrees of freedom ("`DF`") and "`Level of significance`". The symbol "`n.s.`" found to the right of a statistic in other tables means "not significant".

Self-explanatory messages are printed on the screen if difficulties of the following types are encountered: Files cannot be opened, read or closed; an erroneous answer is given during the interactive sequence; limitations on data are exceeded (see Sec. A.2). Messages are printed in the output in the following cases: A requested frequency distribution, measure or test cannot be calculated; differences in definition of genetic types between samples prohibit comparison of the samples; special situations arise during a test. Some messages are followed by "`cause=`" and a number, the latter referring to a compiler-specific list of I/O status values.

5 Frequency distributions

The input to GSED usually consists of the genotypes observed at one or more gene loci in a sample of diploid individuals. (It is also possible to input haplotypes observed in a sample of gametophytes of one sex by listing the second allele at each locus as unknown, *i.e.*, "`-1`", and gametic sex as specified (see Sec. 2.3).) From genotype data, it is possible to construct the following frequency distributions:

For a single locus:

- *Allele frequencies:* Each sampled (diploid) individual contributes two alleles to the overall sample, so that heterozygotes reveal more allelic types than homozygotes. The association between alleles in genotypes (genotypic structure) therefore determines the degree to which a sample detects the allelic types in a population (see Sec. 4: "`alpha`", "`alpha - HWP`").
- *Genotype frequencies:* The genotype of each sampled individual is counted (without regard to gametic sex specification).

Over a set of loci:

- *Multilocus genotype frequencies:* The multilocus genotype of each sampled individual is counted (without regard to gametic sex specification).

Gametic sex specification: In some organisms it is possible to determine which allele at a nuclear gene locus was contributed by the maternal parent. For example, the seed of most coniferous species contains not only the diploid embryo but also nutritive tissue genetically identical to the maternal gametophyte — the primary endosperm or megagametophyte. If the endosperm of a seed is subjected to isoenzyme electrophoresis, the maternal phenotype is revealed. Inheritance analysis of the phenotypes of the endosperm produced by single trees allows inference of the haploid genotype (haplotype) of each endosperm and thus the diploid genotype of each tree ([1]). Inference of the genotype of a diploid embryo and "subtraction" of the haplotype of the corresponding endosperm then reveals the haplotype of the paternal gamete for codominant alleles of enzyme loci ([20], [21]).

If the gametic sex of the alleles (*i.e.*, the sex of the parent contributing each allele) at all involved loci is specified, a number of additional frequency distributions can be calculated:

For a single locus:

- *Allele frequencies among maternal contributions:* The set of alleles contributed by the maternal parents of the sampled individuals represents a sample of the alleles in the "population" of successful maternal gametes.
- *Allele frequencies among paternal contributions:* In like manner, the set of alleles contributed by the paternal parents of the sampled individuals represents a sample of the alleles in the "population" of successful paternal gametes.

- *Ordered genotype frequencies (maternal × paternal alleles):* The set of ordered genotypes represents a sample out of the "population" of successful fusions between female and male gametes. Ordered genotypes take into account the gametic sex specification of the alleles at the locus, distinguishing for example between the genotypes "`1 x 3`" and "`3 x 1`" (see Tab. 9).

Over a set of loci:

- *Haplotype frequencies among maternal contributions:* The set of maternal haplotypes represents a sample of the haplotypes in the "population" of successful maternal gametes.
- *Haplotype frequencies among paternal contributions:* The set of paternal haplotypes represents a sample of the haplotypes in the "population" of successful paternal gametes.
- *Haplotype frequencies:* A sample of the haplotypes of successful gametes is constructed by counting both the maternal and the paternal haplotypes of the sampled individuals. Since each sampled individual contributes two haplotypes, the association between haplotypes in genotypes (genotypic structure) determines the degree to which a sample detects the haplotypes present in a population (see "*Allele frequencies*" above and see Sec. 4: "`alpha`", "`alpha HWP`").
- *Ordered multilocus genotype frequencies (maternal × paternal haplotypes):* The set of ordered genotypes represents a sample of the genotypes in the "population" of successful fusions between female and male gametes. Ordered multilocus genotype frequencies distinguish between maternal and paternal haplotypes. For example, whereas the ordered genotype { `1 2  1 2` } results from fusion of the maternal haplotype { `1 1` } and the paternal haplotype { `2 2` }, the ordered genotype { `2 1  1 2` } is the product of maternal haplotype { `2 1` } and paternal haplotype { `1 2` } (see Tab. 9).

Obtaining unordered genotypes when gametic sex is specified: Note that if gametic sex is specified and the response to the question "`Should gametic sex specification, if given, be retained ?`" is "`Y`" (see Sec. 3.1), then the ordered genotype frequency distribution will be calculated, and all measures will be based on this distribution. In order to obtain the unordered distribution and measures calculated for it, GSED must be restarted using the same input file, but a reply of "`N`" must be given to the above question.

Gene pool: If all of the locus combinations that were chosen for calculation were single loci, then the gene pool made up of the genes at these loci is automatically constructed. This will be the case if option "`0`" or "`1`" was given in answer to "`Locus configuration ?`" of the interactive sequence (see Sec. 3.1). Although the frequency distribution of the gene pool is not explicitly included in the output, all of the chosen measures of variation within

and between samples are also calculated for the gene pool (see Sec. 8) and listed at the end of the output.

Unknown alleles and genotypes: Sometimes it is not possible to determine the genotype or, if gametic sex is specified, one of the parental contributions to an individual at one or more of the investigated loci. In this case, it is up to the user to make sure that the unknown types represent random samples of the respective types in the population. GSED treats unknown alleles (denoted "`-1`" in input) and haplotypes and genotypes containing them as follows for each frequency distribution, :

- *Maternal allele/haplotype frequencies:* Unknown maternal alleles are assumed to be a random sample of all maternal alleles and are thus left out of the calculation. In the same manner, incomplete maternal (multilocus) haplotypes containing an unknown allele at one or more loci are also treated as a random sample of haplotypes and are ignored.
- *Paternal allele/haplotype frequencies:* Unknown paternal alleles and haplotypes are treated in the same way as maternal ones.
- *Allele/haplotype frequencies:* Only those alleles are taken into account that are part of a completely known genotype. Thus if one allele is known and the other is unknown (e.g., because the primary endosperm of a conifer seed was analyzed but the embryo lost), the known allele will not be counted in the allele frequency distribution.
- *Genotype frequencies:* Unknown genotypes are assumed to be a random sample and are not counted.

6 Measures of variation

The following measures of variation can be calculated for any of the types of frequency distributions listed in Sec. 5.

6.1 Measures of variation within samples

6.1.1 Diversity v

Let a collection be characterized by a frequency vector $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ of its genetic types, where $n \in \mathbb{N}$ and for $k = 1, \ldots, n$ $p_k \geq 0$ and $\sum_{k=1}^{n} p_k = 1$. The **diversity** $v(\mathbf{p})$ of the collection is defined as

$$v(\mathbf{p}) = \left(\sum_{k=1}^{n} p_k^2 \right)^{-1}$$

$v(\mathbf{p})$ measures the "differentiation effective number" of types; it is less than or equal to the actual number of types and equals this number only for a uniform distribution.

Reference: [6], [11]

6.1.2 Total population differentiation δ_T

Let a collection of size N be characterized by a frequency vector $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ of its genetic types, where $n \in \mathbb{N}$ and for $k = 1, \ldots, n$, $p_k \geq 0$ and $\sum_{k=1}^{n} p_k = 1$. The **total population differentiation** δ_T of the collection is defined as

$$\delta_T = \frac{N}{N-1} \cdot \left(1 - \sum_{k=1}^{n} p_k^2 \right)$$

or, letting $N_k := N \cdot p_k$ be the absolute frequency of the kth type,

$$\delta_T = \sum_{k=1}^{n} \frac{N_k}{N} \cdot \frac{N - N_k}{N-1}$$

It holds that $0 \leq \delta_T \leq 1$, with $\delta_T = 0$ for monomorphy and $\delta_T = 1$ if no two sample members are of the same genetic type.

References: [11], [12]

6.1.3 Evenness e

"Given a distribution of types of individuals in a collection, the **evenness** of the distribution is considered to measure the degree to which these types are equally represented" [14].

The **evenness** e is defined to equal one minus the minimal distance of the frequency distribution to all "plateaus", each consisting of equally frequent types, in effectively infinite collections. In small collections, the plateaus are defined by the respective distributions closest to uniformity. If $d_{\min}$ equals this minimal distance, the **absolute evenness** is given by

$$e = 1 - d_{\min}$$

(for the definition of d see "genetic distance" below). $e = 1$ holds only for uniform distributions. As e approaches a lower bound of 0.5, the unevenness increases. As a transformation of evenness which varies between 0 and 1, the **relative evenness** of the population is defined as

$$e = 1 - 2 \cdot d_{\min}$$

Reference: [14]

6.2 Measures of variation between samples

6.2.1 Genetic distance d_0

Let two collections be characterized by frequency vectors $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ and $\mathbf{p}' = (p'_1, p'_2 \ldots, p'_n)$ of their genetic types, where $n \in \mathbb{N}$ and for $k = 1, \ldots, n$, $p_k, p'_k \geq 0$ and $\sum_{k=1}^{n} p_k = 1 = \sum_{k=1}^{n} p'_k$. The **genetic distance** $d_0(\mathbf{p}, \mathbf{p}')$ is defined as

$$d_0(\mathbf{p}, \mathbf{p}') = \frac{1}{2} \cdot \sum_{k=1}^{n} |p_k - p'_k|$$

The genetic distance between two collections is specified as the proportion of genetic elements (alleles, genes at multiple loci, gametes, genotypes) which the two collections do not share. Thus $d_0 = 1$ if and only if the two collections have no types in common.

References: [4], [5], [6], [8]

6.2.2 Subpopulation differentiation D_j and δ

Let a population be divided into demes (subpopulations, collections). The amount of **genetic differentiation** of one subpopulation to the remainder of the population is specified as "the proportion of genetic elements (alleles, genes at multiple loci, gametes, genotypes) by which a deme differs from the remainder of the population in type" [9]. This proportion is defined as

$$D_j = d_0(\mathbf{p}_j, \bar{\mathbf{p}}_j)$$

where $\mathbf{p}_j$ and $\bar{\mathbf{p}}_j$ are the frequency distributions of the types in deme j and in the remainder of the population, respectively, and d_0 is the genetic distance defined above. The **subpopulation differentiation** is then defined by

$$\delta = \sum_j c_j \cdot D_j$$

where the weights c_j express the proportion of genetic elements present in the jth deme.

References: [9], [10], [12], [15]

6.2.3 Test of homogeneity

Let m collections of individuals each be characterized by a frequency distribution defined by the number of individuals of each of n types in the collection. A **test of homogeneity** of the m frequency distributions tests the hypothesis that these m collections all originated from a single large collection of individuals, conditioned on the marginal distributions given by the m sample sizes as proportions of the sum of sample sizes and the mean relative frequencies of the n types over the samples. Goodness-of-fit tests (see Sec. 7.2) are performed for $(m-1)(n-1)$ degrees of freedom.

References: [2] pp.365*ff*, [24] pp.96*ff*

7 Analysis of genotypic structure

The following measures and tests aid in the characterization of genotypic structures. In contrast to other measures quantifying variation within and between samples (see Sec. 6, 8), heterozygosity measures genetic variation within individuals. Tests of single locus structure investigate the association of gametes in observed (zygotic) genotypic structures by comparing the observed structures to the corresponding expected structures under certain models of association.

7.1 Heterozygosity

7.1.1 Proportion of heterozygosity of single-locus genotypes

Given the genotypes of all individuals in a collection at a single gene locus, the **proportion of heterozygosity** equals the proportion of heterozygous individuals in the collection.

Reference: [16]

7.1.2 Conditional heterozygosity of single-locus genotypes

The **conditional heterozygosity** at a single gene locus takes into account that the proportion of heterozygosity is conditional on the allele frequencies. It results from division of the actual heterozygosity (= proportion of heterozygosity at a single locus) by the corresponding maximum proportion of heterozygosity $H_{\max}$ obtainable for the underlying allele frequencies, where $H_{\max}$ equals 1 if all allele frequencies are less than or equal to 0.5 and $H_{\max} = 2(1 - p)$ if the most frequent allele has frequency p greater than 0.5.

References: [6], [16]

7.1.3 Degree of heterozygosity of multilocus genotypes

The **degree of heterozygosity** is defined for an individual with respect to a specified number of gene loci, and is identical to the proportion of loci at which this individual is heterozygous. The **average degree of heterozygosity** refers to the distribution of this degree in a collection of individuals. Hence it can be proven that the **average degree of heterozygosity** equals the mean proportion of heterozygotes at the single loci.

Reference: [6]

7.2 Tests of single locus structure

The following goodness-of-fit tests are performed for two models of single locus genotypic structure: Pearson's χ^2 **goodness-of-fit test** with statistic

$$X^2 = \sum_{\text{types}} \frac{(N_\cdot - E(N_\cdot))^2}{E(N_\cdot)}$$

Likelihood ratio test with statistic

$$G = 2 \cdot \sum_{\text{types}} N_\cdot \cdot (\ln(N_\cdot) - \ln(E(N_\cdot)))$$

For one degree of freedom, χ^2 **goodness-of-fit test with continuity correction c** $= \frac{1}{2}$ with statistic

$$X^2_{c=\frac{1}{2}} = \sum_{\text{types}} \frac{\left(|N_\cdot - E(N_\cdot)| - \frac{1}{2}\right)^2}{E(N_\cdot)}$$

$N_\cdot$ and $E(N_\cdot)$ represent observed and expected sample frequencies, respectively, of the different types.

These statistics are asymptotically χ^2-distributed, the number of degrees of freedom depending on the model. Thus it must be kept in mind that these tests are accurate only for large sample sizes. (A warning is printed in the output if a type is found to have expected frequency less than 5.)

Exact tests have recently been devised in some cases, but these seemed too time-consuming in terms of computing time to allow their inclusion in the larger framework of GSED. In borderline cases (*i.e.*, statistic near critical value of χ^2) of small sample size, it may be advisable to retest structures using special statistics programs that perform exact hypothesis testing.

References: [19], [25] pp.71*ff*

7.2.1 Test of Hardy-Weinberg structure and heterozygosity

To each unordered genotypic structure with relative frequencies P_{ij} of genotypes A_iA_j ($P_{ij} = P_{ji}$, $\sum_{i \le j} Pij = 1$) there corresponds a Hardy-Weinberg structure with genotypic frequencies P^*_{ij} defined by

$$P^*_{ii} = p_i^2 \quad \text{and} \quad P^*_{ij} = 2p_ip_j \quad \text{for } i \neq j \text{ and } i, j = 1, \ldots, k$$

In this definition, p_i is the relative frequency of allele A_i from the original genotypic structure, *i.e.*, $p_i = P_{ii} + \frac{1}{2}\sum_{j,j\neq i} P_{ij}$. Hardy-Weinberg structures result from special mating systems, such as are specified, e.g., in [13] pp. 20*ff*, 68*ff*, and [17] pp. 175*ff*.

The purpose here is to detect deviations of (1) an actual genotypic structure (P_{ij}) from its corresponding Hardy-Weinberg structure (P^*_{ij}) and (2) actual heterozygosity from the

corresponding Hardy-Weinberg heterozygosity. Actual heterozygosity is defined by $P_{\text{het}} = 1 - \sum_i P_{ii}$ and its corresponding Hardy-Weinberg heterozygosity by $P^*_{\text{het}} = 1 - \sum_i p_i^2$.

Assume that a sample of N individuals was randomly drawn from a large population, and consider their genotypes at a locus with k alleles. (Gametic sex, if specified, is disregarded, *i.e.*, genotypes A_iA_j and A_jA_i are not distinguished.) For unordered absolute genotype frequencies N_{ij} $(i, j = 1, \ldots, k,\ N_{ij} = N_{ji},\ \sum_{i \le j} N_{ij} = N)$ in the sample, the absolute frequency N_i of allele A_i in the sample of $2N$ alleles equals $N_i = 2N_{ii} + \sum_{j \ne i} N_{ij}$. Conditioning on the allele frequencies in the sample (*i.e.*, assuming that the true frequency p_i of allele A_i in the population equals $p_i = N_i/(2N)$), the genotypic frequencies expected under the null hypothesis of Hardy-Weinberg structure equal

$$E(N_{ii}) = N_i^2/(4N) \quad \text{and} \quad E(N_{ij}) = N_iN_j/(2N) \quad (i, j = 1, \ldots, k)$$

The N_{ij} and $E(N_{ij})$ for $i \le j$ are the observed and expected sample frequencies, respectively, entering the test statistics described above. The number of degrees of freedom equals $k(k-1)/2$.

The observed numbers of homozygotes and heterozygotes in a sample of N individuals from a large population equal $\sum_i N_{ii}$ and $\sum_{i<j} N_{ij}$, respectively. The numbers of homozygotes and heterozygotes expected under the assumption of a Hardy-Weinberg structure equal

$$E\left(\sum_i N_{ii}\right) = \left(\sum_i N_i^2\right)/(4N) \quad \text{and} \quad E\left(\sum_{i<j} N_{ij}\right) = N - E\left(\sum_i N_{ii}\right)$$

respectively. Again, the expected frequencies are conditioned on the allele frequencies in the sample. One degree of freedom remains.

By definition, a genotypic structure shows an excess of homozygotes (heterozygotes) if its proportion of homozygotes (heterozygotes) exceeds the proportion of homozygotes (heterozygotes) in the corresponding Hardy-Weinberg structure. If the genotypic structure is a Hardy-Weinberg structure, then such an excess will not be statistically significant; if the test for Hardy-Weinberg structure is not significant, an excess still may or may not be significant. Tests for homozygote excess frequently form the first step in an analysis of so-called "inbreeding structures". Detailed tests for realization of inbreeding structures require consideration of various cases as specified, e.g., in [23].

References: [2], [13] pp.20*ff*, 68*ff*, [17] pp.175*ff*, [18], [22], [23] and references therein, [25] pp.71*ff*

7.2.2 Test of product structure for ordered genotypes

In a large population, random fusion of gametes from the set of maternal and the set of paternal gametes gives rise to a zygotic genotypic structure at a locus with k alleles that fulfills the properties of a product structure

$$P_{ij} = p_i^{\female} p_j^{\male} \quad (i, j = 1, \ldots, k)$$

where P_{ij} is the relative frequency of the ordered genotype A_iA_j (*i.e.*, A_i is the maternal contribution and A_j the paternal, so that $\sum_{i,j} P_{ij} = 1$), $p_i^{\female}$ is the relative frequency of allele A_i among maternal gametic contributions, and $p_i^{\male}$ is the relative frequency of allele A_i among paternal gametic contributions.

Given a random sample of N individuals from a large population, the test of a product structure is performed as a test of independence of association between maternal and paternal allelic contributions conditioned on marginal distributions given by the frequencies of these contributions in the sample. For absolute frequencies N_{ij} $(i, j = 1, \ldots, k)$ of the ordered genotypes in the sample, the absolute frequency $N_i^{\female}$ of the allele A_i in the sample of N maternal alleles equals $N_i^{\female} = \sum_{j=1}^{k} N_{ij}$ and the frequency $N_i^{\male}$ of the allele A_i in the sample of N paternal alleles equals $N_i^{\male} = \sum_{j=1}^{k} N_{ji}$. Conditioning on the allele frequencies in the sample (*i.e.*, assuming that the true frequency $p_i^{\female}$ and $p_i^{\male}$ of allele A_i among the maternal and paternal gametes produced in the population equals $p_i^{\female} = N_i^{\female}/N$ and $p_i^{\male} = N_i^{\male}/N$, respectively), the genotypic frequencies expected under the null hypothesis of a product structure equal

$$E(N_{ij}) = N_i^{\female} N_i^{\male}/N \quad (i, j = 1, \ldots, k)$$

The number of degrees of freedom equals $(k^{\female} - 1)(k^{\male} - 1)$, where $k^{\female}$ and $k^{\male}$ are the numbers of alleles with non-zero frequency among maternal and paternal contributions, respectively.

Reference: [2] pp.360*ff*

8 Analysis of the gene pool

"The gene pool of a population with respect to the number [L] of (non-homologous) gene loci located at a certain section of the genome is thought of as the set of all gene (alleles) at these loci realized in all individuals" [6]

The following types of gene pool can be constructed, the first two only if gametic sex is specified at all contributing loci:

- Gene pool of maternal contributions
- Gene pool of paternal contributions
- Gene pool of single-locus genotypes

8.1 Measures of variation within samples

8.1.1 Diversity v of the gene pool

Let a collection be characterized at each of L loci by the frequency vector $\mathbf{p}_l = (p_{1l}, p_{2l}, \ldots, p_{nl})$ for $l = 1, \ldots, L$, where $n \in \mathbb{N}$ and for $i = 1, \ldots, n$, $p_{il} \geq 0$ and $\sum_{i=1}^{n} p_{il} = 1$. Denoting by

$$v_{(l)} = \left(\sum_{i=1}^{n} p_{il}^2 \right)^{-1}$$

the allelic diversity at the l-th locus, the **gene pool (genic) diversity** v of the collection was proved to equal the harmonic mean of the single-locus diversities, *i.e.*,

$$v = \left(\frac{1}{L} \cdot \sum_{l=1}^{L} \frac{1}{v_{(l)}} \right)^{-1} = \frac{L}{\sum_{l=1}^{L} \left(\sum_{i=1}^{n} p_{il}^2 \right)}$$

Reference: [11]

8.1.2 Diversity v_{gam} of the hypothetical gametic output

Let a collection be characterized at locus l ($l = 1, \ldots, L$) by the frequency vector $\mathbf{p}_l = (p_{1l}, p_{2l}, \ldots, p_{nl})$, where $n \in \mathbb{N}$ and for $i = 1, \ldots, n$, $p_{il} \geq 0$ and $\sum_{i=1}^{n} p_{il} = 1$. Denoting by

$$v_{(l)} = \left(\sum_{i=1}^{n} p_{il}^2 \right)^{-1}$$

the allelic diversity at the l-th locus, the **hypothetical gametic diversity** v_{gam} of the collection is defined as

$$v_{gam} = \prod_{l=1}^{L} v_{(l)}$$

The hypothetical gametic output is defined by the set of gametes that results from stochastically independent association between loci, free recombination, and equal gametic production for all members. v_{gam} therefore measures the potential of a population for producing genetically diverse gametes.

Reference: [6]

8.1.3 Total population differentiation δ_T of the gene pool

Let a collection of subpopulations have the total population differentiation $\delta_{T(l)}$ at locus l ($l = 1, \ldots, L$). Then the **total population differentiation δ_T of the gene pool** was proven to equal the arithmetic mean of the total population differentiation at each locus, that is,

$$\delta_T = \frac{1}{L} \cdot \sum_{l=1}^{L} \delta_{T(l)}$$

Reference: [11]

8.2 Measures of variation between samples

8.2.1 Distance d_0 between gene pools

Let one collection be characterized by the frequency vectors of the different genes (alleles) at L gene loci, that is, by the L frequency vectors $\mathbf{p}_l = (p_{1l}, p_{2l}, \ldots, p_{n_l l})$ $(l = 1, \ldots, L)$, where $n_l \in \mathbb{N}$ is the number of alleles at locus l and $p_{kl} \geq 0$ and $\sum_{k=1}^{n_l} p_{kl} = 1$ holds for all $k = 1, \ldots, n_l$. Let a second collection be characterized by the L frequency vectors $\mathbf{p}'_l = \left(p'_{1l}, p'_{2l}, \ldots, p'_{n_l l}\right)$ $(l = 1, \ldots, L)$ at the same L loci and for the same numbering of alleles at each locus. The **gene pool genetic distance** d_0 between the two collections was proven to be the arithmetic mean of the single-locus distances, *i.e.*,

$$\begin{aligned} d_0 &= \frac{1}{L} \cdot \sum_{l=1}^{L} d_0(\mathbf{p}_l, \mathbf{p}'_l) \\ &= \frac{1}{L} \cdot \sum_{l=1}^{L} \left(\frac{1}{2} \cdot \sum_{k=1}^{n_l} |p_{kl} - p'_{kl}| \right) \end{aligned}$$

Reference: [10]

8.2.2 Differentiation δ of subdivided gene pools

Let a collection of subpopulations have the subpopulation differentiation $\delta_{(l)}$ at locus l $(l = 1, \ldots, L)$. Then the (unweighted) **subpopulation differentiation δ of the gene pool** was proven to be the arithmetic mean of the subpopulation differentiation at each locus, that is

$$\delta = \frac{1}{L} \cdot \sum_{l=1}^{L} \delta_{(l)}$$

Reference: [10]

9 Acknowledgements and disclaimer

I am extremely grateful to the many colleagues who have been of enormous help in completing this project. Matthias Kühle did some of the programming of the interactive sequence, the configuration file being his idea. His ability to find his way around in somebody else's program, which at that time was already complex, is admirable. I would like to thank numerous colleagues, among them Fritz Bergmann, Bernd Degen, Reiner Finkeldey, Hans-Rolf Gregorius, Hans H. Hattemer, Sven Herzog, Bernhard Hosius, Gerhard Müller-Starck, Aristotelis Papageorgiou, Rommy Starke, Jozef Turok, Martin Ziehe, and too many master's students to list here for their eagerness to test the various stages of the program on their data and for their many suggestions for improvement. Martin Ziehe, in particular, is to be credited for taking the trouble to recalculate many of the computed results, thereby discovering a number of bugs before they could do any damage. Special thanks are due to Hans-Rolf Gregorius for his valuable instruction over the years on the meaning of genetic variation and mating systems in general and on the implemented measures and tests in particular as well as for concrete suggestions for improvement of the output. Finally, I am grateful to Hans-Rolf Gregorius, Hans H. Hattemer, Bernhard Hosius, and Martin Ziehe for helpful suggestions for improvement of this manual.

Generous financial support, without which this project could not have been completed so soon, was provided by two institutions: the Bundesforschungsanstalt für Forst- und Holzwirtschaft (BFH) in Hamburg and the Hessische Forstliche Versuchsanstalt in Hann. Münden. Florian Scholz and Bernd Degen of the Institut für Forstgenetik, BFH, Großhansdorf, and Alwin Janßen of the Hess. Forstl. Versuchsanstalt were instrumental in obtaining this support, and I thank them for their very considerable efforts.

I have tried my best to find all programming errors. Nevertheless, the user is advised to check the correctness of the results, as I can assume no liability for any errors I've missed. I would be very grateful for news of any real mistakes that remain in the program.

References

[1] Bergmann, F. 1971. Genetische Untersuchungen bei *Picea abies* mit Hilfe der Isoenzym-Identifizierung. II. Möglichkeiten für genetische Zertifizierung von Forstsaatgut. *Allgemeine Forst- und Jagdzeitung* 142, 278–280.

[2] Elandt-Johnson, R.C. 1971. *Probability Models and Statistical Methods in Genetics.* John Wiley & Sons, Inc., New York, London, Sydney, Toronto.

[3] Emigh, T.H. 1980. A comparison of tests for Hardy-Weinberg equilibrium. *Biometrics* 36, 627–642.

[4] Gregorius, H.-R. 1974. Genetischer Abstand zwischen Populationen. I. Zur Konzeption der genetischen Abstandsmessung. *Silvae Genetica* 23, 22–27.

[5] Gregorius, H.-R. 1974. On the concept of genetic distance between populations based on gene frequencies. *Proc. Joint IUFRO Meeting, S02.04.1-3, Stockholm, Session I*, 17–26.

[6] Gregorius, H.-R. 1978. The concept of genetic diversity and its formal relationship to heterozygosity and genetic distance. *Math. Biosciences* 41, 253–271.

[7] Gregorius, H.-R. 1980. The probability of losing an allele when diploid genotypes are sampled. *Biometrics* 36, 643–652.

[8] Gregorius, H.-R. 1984. A unique genetic distance. *Biometrical J.* 26, 13–18.

[9] Gregorius, H.-R. 1984. Measurement of genetic differentiation in plant populations. Pp. 276–285 in: Gregorius, H.-R. (ed.). *Population Genetics in Forestry.* Springer-Verlag, Berlin, Heidelberg, New York, Tokyo.

[10] Gregorius, H.-R. and Roberds, J.H. 1986. Measurement of genetical differentiation among subpopulations. *Theor. Appl. Genet.* 71, 826–834.

[11] Gregorius, H.-R. 1987. The relationship between the concepts of genetic diversity and differentiation. *Theor. Appl. Genet.* 74, 397–401.

[12] Gregorius, H.-R. 1988. The meaning of genetic variation within and between subpopulations. *Theor. Appl. Genet.* 76, 947–951.

[13] Gregorius, H.-R. 1989. *Characterization and Analysis of Mating Systems.* Ekopan Verlag, Witzenhausen.

[14] Gregorius, H.-R. 1990. A diversity-independent measure of evenness. *Amer. Natur.* 136, 701–711.

[15] Gregorius H-R (1996) Differentiation between populations and its measurement. *Acta Biotheoretica* 44, 23–36.

[16] Gregorius, H.-R., Krauhausen, J., and Müller-Starck, G. 1986. Spatial and temporal genetic differentiation among the seed in a stand of *Fagus sylvatica* L. *Heredity* 57, 255–262.

[17] Hattemer, H.H, Bergmann, F., Ziehe, M. 1993. *Einführung in die Genetik für Studierende der Forstwissenschaft*, 2. Aufl. J.D. Sauerländer's Verlag, Frankfurt am Main.

[18] Ledwina, T., and Gnot, S. 1980. Testing for Hardy-Weinberg equilibrium. *Biometrics* 36, 161–165.

[19] Louis, E.J., and Dempster, E.R. 1987. An exact test for Hardy-Weinberg and multiple alleles. *Biometrics* 43, 805–811.

[20] Müller[-Starck], G. 1977. Untersuchungen über die natürliche Selbstbefruchtung in Beständen der Fichte (*Picea abies* (L.) Karst.) und Kiefer (*Pinus sylvestris* L.). *Silvae Genetica* 26, 207–217.

[21] Müller[-Starck], G. 1977. Cross-fertilization in a conifer stand inferred from enzyme gene markers in seeds. *Silvae Genetica* 26, 223–226.

[22] Pamilo, P., and Varvio-Aho, S. 1984. Testing genotype frequencies and heterozygosities. *Marine Biology* 79, 99–100.

[23] Robertson, A., and Hill, W.G. 1984. Deviations from Hardy-Weinberg proportions: Sampling variances and use in estimation of inbreeding coefficients. *Genetics* 107, 703–718.

[24] Weber, E. 1978. *Mathematische Grundlagen der Genetik.* VEB Gustav Fischer Verlag, Jena.

[25] Weir, B.S. 1990. *Genetic Data Analysis.* Sinauer Associates, Inc. Publ., Sunderland, Mass.

A Technical considerations

A.1 Hardware requirements

GSED is written in the programming language FORTRAN-77 and can be implemented under any operating system for which a FORTRAN-77 compiler is available. To date it has been compiled for the operating systems MS-DOS/PC-DOS (Version 2.1 or higher), VMS (Version 5.0 or higher), and TOS (Atari). A math co-processor is not used.

GSED spends most of its time writing intermediate results to direct access files and retrieving them from these (see Sec. A.3). Since these files are stored in the default directory (VMS, TOS) or in the root directory of the default drive (DOS), this process can be speeded up considerably if the executable program is started from a RAM-Disk (RAM = Random-Access-Memory) as the default drive. The executable program itself may be read from a different device (e.g., floppy disk), as long as the prompt points to the RAM-Disk as the default drive. Locating the input and output files in the RAM-Disk additionally speeds up calculation. (Moreover, operating in a RAM-Disk prevents lots of wear and tear on the hard disk drive.) A 1 MB RAM-Disk should be large enough for most applications. Except for very small input files, it is not recommended to use a floppy disk as the default drive, since calculation is extremely slow.

A.2 Limitations on data

Tab. 11 gives a list of variables, the maximal allowed values of which were fixed at compile-time and can be changed only by altering the source code. Error messages are printed if any of these maxima is exceeded, and execution is stopped.

The test of homogeneity allows a maximum of *Max.No.Alleles* genetic types. All tests accomodate up to 100 degrees of freedom. Violation does not cause termination of execution in these cases.

All integers are of type INTEGER*4 and range between -2147483647 and +2147483647. Reals are of single precision type REAL*4 with approximately 7 digit accuracy and range from 10^{-38} to 10^{38}.

The output formats accomodate 5-digit integers (up to 99999) and floating point numbers with up to 5 digits in front of the decimal point. Floating point calculations are printed with 3 decimal places; the one exception are the expected absolute frequencies in tests, which have two decimal places.

A.3 Temporary direct access storage

GSED stores intermediate results in eight direct access files. Depending on the compiler they receive names such as “FORnnn.DAT”, “F77$nnnn.$$$”, or a seemingly arbitrary sequence of letters and numbers. They are stored on the default directory or drive (see

Table 11: *Interpretation and maximal values of variables used in calculating storage space needed for direct access files*

Variable	Maximum	Interpretation
No.Samples	100	Number of samples in input file
No.Chosen.Samples		Number of chosen samples (see Sec. 3.1: "`How many samples ?`")
Max.No.Identified		Maximum number of individuals identified per sample (see Sec. 4)
No.Loci	50	Number of gene loci in input data
No.Locus.Combinations	60	Number of locus combinations (see Sec. 3.1: "`Locus configuration ?`"
Max.No.Loci.Per.Combination		Maximum number of loci per locus combination (=1 if single locus only, see Sec. 3.1: "`Locus configuration ?`")
Max.No.Alleles	40	Maximum number of different alleles at any locus and in any sample
Max.No.Haplotypes	500	Maximum number of different haplotypes that occur for any multilocus combination and any sample (see Sec. A.3)
Max.No.Genotypes	500	Maximum number of different genotypes that occur for any locus combination and sample
Size.of.Enlargement		Size of storage space added when file is enlarged, expressed as the equivalent number of 4-byte integers

Table 12: *Maximum size of the direct access files as number of 4-byte integers. Multiplication of each result by 4 yields total number of bytes. Notation: (1) If additional storage space is added blockwise, add Size.of.Enlargement. (2) Max.No.Haplotypes = Max.No.Alleles if only single loci are considered.*

Unit	File size as number of 4-byte integers[(1)]
20	$(Max.No.Haplotypes^{(2)} + 1) * No.Locus.Combinations * No.Samples * 3$
21	$(Max.No.Haplotypes^{(2)} + 1) * No.Locus.Combinations * Max.No.Loci.Per.Combination$
22	$(Max.No.Genotypes + 1) * No.Locus.Combinations * No.Samples$
23	$(Max.No.Genotypes + 1) * No.Locus.Combinations * 2$
24	$(Max.No.Haplotypes^{(2)} + 1) * No.Locus.Combinations$
25	$(Max.No.Genotypes + 1) * No.Locus.Combinations$
26	$(Max.No.Loci.Per.Combination + 2) * No.Locus.Combinations * No.Samples - 1$
27	$Max.No.Identified * No.Chosen.Samples * 2$

Sec. A.1) and are deleted automatically upon successful completion of the program. If the program is interrupted in mid-run, the files may remain and can be deleted by hand. Since they are not in ASCII code, they cannot be read by a text editor.

Sufficient storage space must therefore be available not only for the executable program, the input file, and the output file, but also for the direct access files. The maximum size of these files can be calculated by the formulae given in Tab. 12, which use the variables defined in Tab. 11.

The total storage space for the eight files as number of 4-byte integers is less than or equal to

$$
\begin{aligned}
& (Max.No.Haplotypes + 1) * No.Locus.Combinations * \\
& *(3 * No.Samples + Max.No.Loci.Per.Combination + 1) \; + \\
+ \quad & (Max.No.Genotypes + 1) * No.Locus.Combinations * (No.Samples + 3) \; + \\
+ \quad & (Max.No.Loci.Per.Combination + 2) * \\
& *No.Locus.Combinations * No.Samples + \\
+ \quad & 2 * Max.No.Identified * No.Chosen.Samples + 8 * Size.of.Enlargement
\end{aligned}
$$

where *Max.No.Haplotypes* is set to *Max.No.Alleles* if all locus combinations refer to single loci. Examples follow:

Sample calculations of storage space requirements for direct access files — single-locus genotypes: An input file consists of 10-locus unordered genotypes of 100 individuals in each of 10 samples. A maximum of 4 different alleles appears at each locus, so that at

most 10 genotypes can be constructed at each locus. The values of the variables are shown in the following list. The total storage space required for all eight direct access files equals 5330 integers. For 4-byte integers, the total storage space is 21328 bytes (< 22 kilobytes).

Max.No.Haplotypes=4	*Max.No.Genotypes*=10
No.Locus.Combinations=10	*No.Samples*=10
Max.No.Loci.Per.Combination=1	*Max.No.Identified*=100
No.Chosen.Samples=10	*Size.of.Enlargement*=0

Sample calculations of storage space requirements for direct access files — multilocus genotypes: For the data described in the previous example, the multilocus genotypes defined by all 10 loci are considered. If all individuals in any given sample have different 10-locus genotypes, then the values of the variables are as shown in the following list, and (7574 integers) = (30296 bytes) < (31 kilobytes) storage space are needed.

Max.No.Haplotypes=100	*Max.No.Genotypes*=100
No.Locus.Combinations=1	*No.Samples*=10
Max.No.Loci.Per.Combination=10	*Max.No.Identified*=100
No.Chosen.Samples=10	*Size.of.Enlargement*=0

A note on "Max.No.Haplotypes": In some cases this variable has a biological meaning, while in others it is a mere construct for defining storage space. The reason is that all genotypes are regarded as the product of "haplotypes", which themselves are numbered consecutively in the order in which they are encountered in the input. This manner of storage is the most space-saving, since storage space is used only for genetic types that actually have been found and thus need not be reserved for potential types that never appear. "Haplotypes" in this sense are defined as follows:

1. If all locus combinations specify single loci, *Max.No.Haplotypes* is set to *Max.No.Alleles*, as mentioned above. Each genotype is stored as the product of its two alleles.

2. In the case of locus combinations specifying ordered multilocus genotypes, which are composed of two known haplotypes contributed by the parents, *Max.No.Haplotypes* is the maximum number of different haplotypes (regardless of gametic sex) encountered for any locus combination and in any sample. Each genotype is stored as the product of its two haplotypes.

3. In the case of multilocus combinations lacking gametic sex specification, "haplotypes" are formed artificially as follows: Before a multilocus genotype is stored, the alleles at each of the constituent single loci are put in nondecreasing order (e.g., genotype `2 1` becomes `1 2`). One "haplotype" is then formed as the list of the alleles appearing in the first position at each constituent locus; the second "haplotype" is formed accordingly. *Max.No.Haplotypes* is thus the maximal number of

such "haplotypes" for any locus combination and in any sample. Finally, if a mixture of single locus and ordered and unordered multilocus combinations is chosen, *Max.No.Haplotypes* is the maximum over all of these values.

B Examples of input files

Example 1: The file EXAMPLE.DAT was already introduced (see Sec. 2, 3, and especially Tab. 5).
The first sample ("`Population 1`") consists of the 4-locus genotypes of 5 individuals designated "`1`", "`3`", "`4`", "`5`", and "`6`"; gametic sex is specified at all loci.
"`Population 2`" consists of the 4-locus genotypes of the 5 individuals "`1`"–"`5`"; gametic sex is specified at loci 1 and 2 but not loci 3 and 4.
In "`Population 3`", nine different genotypes were found in a sample of 100 individuals, the frequencies of the different genotypes in the sample equalling 32, 19, ...; gametic sex is not specified at any locus. This constellation of gametic sex specification may not be very realistic, but it demonstrates the form of data input and in particular the meaning of the key line (see Sec. 2.3.1).

```
   3   4
LAP-A
LAP-B
IDH
PGI
Population 1
Population 2
Population 3
 4  (2i4,4(1x,2i2))
   1   0  1 1  2 1  3 1  4 1
   1   1  1 3  1 2  3 3  2 3
   1   3  3 1  1 1  3 3  2 2
   1   4  3 1  1 1  3 3  2 2
   1   5  1 1  1 1  3 3  2 2
   1   6  3 3  2 1  1 3  3 3
9999
   2   0  1 1  2 1  3 0  4 0
   2   1  3 3  1 2  1 1  2 3
   2   2  1 3  2 1  1 3  2 3
   2   3  3 1  1 1  3 3  2 3
   2   4  1 1  1 1  1 3  2 2
   2   5  3 3  1 2  3 3  2 3
9999
   3   0 -1 0  2 0  3 0  4 0
   3  32  3 3  1 2  1 1  2 3
   3  19  1 3  1 2  1 3  2 3
   3   4  1 3  1 1  3 3  2 3
   3   7  1 1  1 1  1 3  2 2
   3  25  3 3  1 2  3 3  2 3
   3   4  1 3  1 1  1 3  2 3
   3   3  3 3  1 2  3 3  2 3
   3   3  1 3  1 1  1 3  3 3
   3   3  1 3  1 2  1 3  2 3
9999
```

Example 2: This example demonstrates that it is still possible to input single locus genotypes at more than one locus, even if they refer to the same population but not necessarily the same individuals. The disadvantage of this type of input is that all individual multilocus information is lost. The "`-1`" in the key lines indicates that the second field of each genotype line contains the frequency with which the respective genotype was found in the sample. Note that several key lines per sample are possible. (Adapted from Kim, Z.-S. 1985. Viability selection at an allozyme locus during development in European beech (*Fagus sylvatica* L.) *Silvae Genetica* 34: 181–186.)

```
   2   2
SAP-A
LAP-A
DEUTSCHLAND: ECKERN
DEUTSCHLAND: KEIMLINGE GEWAECHSHAUS
 1  (2I4,1X,1(2I2,1X))
   1   0 -2 0
   1 151  1 1
   1 111  2 2
   1 107  3 3
   1   6  4 4
   1  51  1 2
   1  83  1 3
   1   5  1 4
   1  68  2 3
   1   7  2 4
   1   3  3 4
9999
   2   0 -1 0
   2  23  1 1
   2  71  2 2
   2   1  3 3
   2  62  1 2
   2   2  1 3
   2   3  1 4
   2   6  2 3
   2   6  2 4
   2   0 -2 0
   2  39  1 1
   2  53  2 2
   2  32  3 3
   2   2  4 4
   2   6  1 2
   2   9  1 3
   2   1  1 4
   2  16  2 3
   2   2  2 4
   2   2  3 4
9999
```

Example 3: Here the number of loci is large, necessitating continuation of the genotypes on additional lines. Note that since all alleles are known (*i.e.*, $\neq$ "`-1`"), the format specification has been altered to leave out the blank between the alleles at a single locus. Lack of gametic sex specification is indicated by the "`0`" following each locus number in the key line.

```
   1  30
Locus  1
Locus  2
...
Locus 30
Beech forest
30  (2i4,20(i2,i1)/8x,10(i2,i1))
   1   0 10 20 30 40 50 60 70 80 90100110120130140150160170180180200
        210220230240250260270280290300
   1   1 11 22 31 11 21 12 44 34 35 43 31 32 33 22 13 13 22 11 11 22
         43 23 22 33 11 22 31 12 11 00
   1   2 22 33 33 22 44 11 22 11 31 11 22 31 12 14 34 42 13 23 21 22
         35 43 21 12 11 23 21 11 22 12
9999
```

Example 4: This example shows the input for a sample of "successful" maternal haplotypes, such as could be found by sampling the bulk seed of a stand of a conifer species and subjecting only the primary endosperm of each seed to isoenzyme electrophoresis (see Sec. 5). The specification of gametic sex is indicated by the "`1`" following each locus number in the key line. The unknown paternal contribution at each locus is designated by "`-1`".

```
   1   4
Locus 1
Locus 2
Locus 3
Locus 4
Scots pine forest
 4  (2i4,4(i3,i2))
   1   0  1 1  2 1  3 1  4 1
   1   1  2-1  3-1  3-1  1-1
   1   2  2-1  2-1  1-1  2-1
   1   3  1-1  3-1  2-1  3-1
   1   4  1-1  2-1  1-1  2-1
9999
```